Linear Regression

Data: insurance.csv

Insurance Company: premium > claims

Structure of data

```
> insurance <- read.csv("C:/Users/Dr Vinod/Desktop/insurance.csv")
>   View(insurance)
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ..
.
 $ charges : num  16885 1726 4449 21984 3867 ...
```

Out of 7 variables, 3 are Factors, 2 Integers & 2 Numeric

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	0	no	northwest	28923.137
11	25	male	26.220	0	no	northeast	2721.321
12	62	female	26.290	0	yes	southeast	27808.725

Description of variables

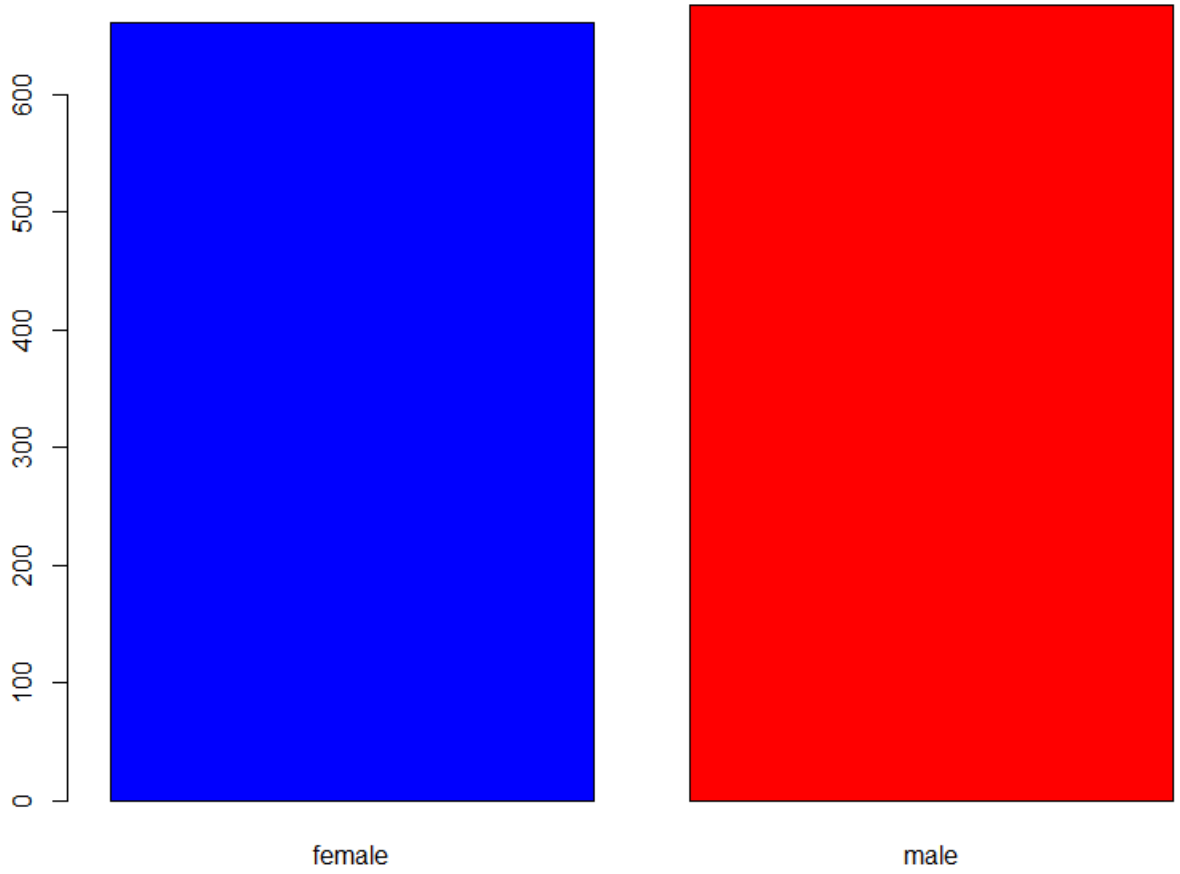
- `age`: An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
- `sex`: The policy holder's gender, either male or female.
- `bmi`: The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
- `children`: An integer indicating the number of children/dependents covered by the insurance plan.
- `smoker`: A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
- `region`: The beneficiary's place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.

Females and Males in data

Almost same proportion

```
b = table(insurance$sex)

b

barplot(b, col = c("blue", "red"))
```

```
> b

female   male
   662    676
```



```
c = table(insurance$children)

c

barplot(c, col = c("blue", "red", "yellow",
                   "green", "purple", "pink"))
```

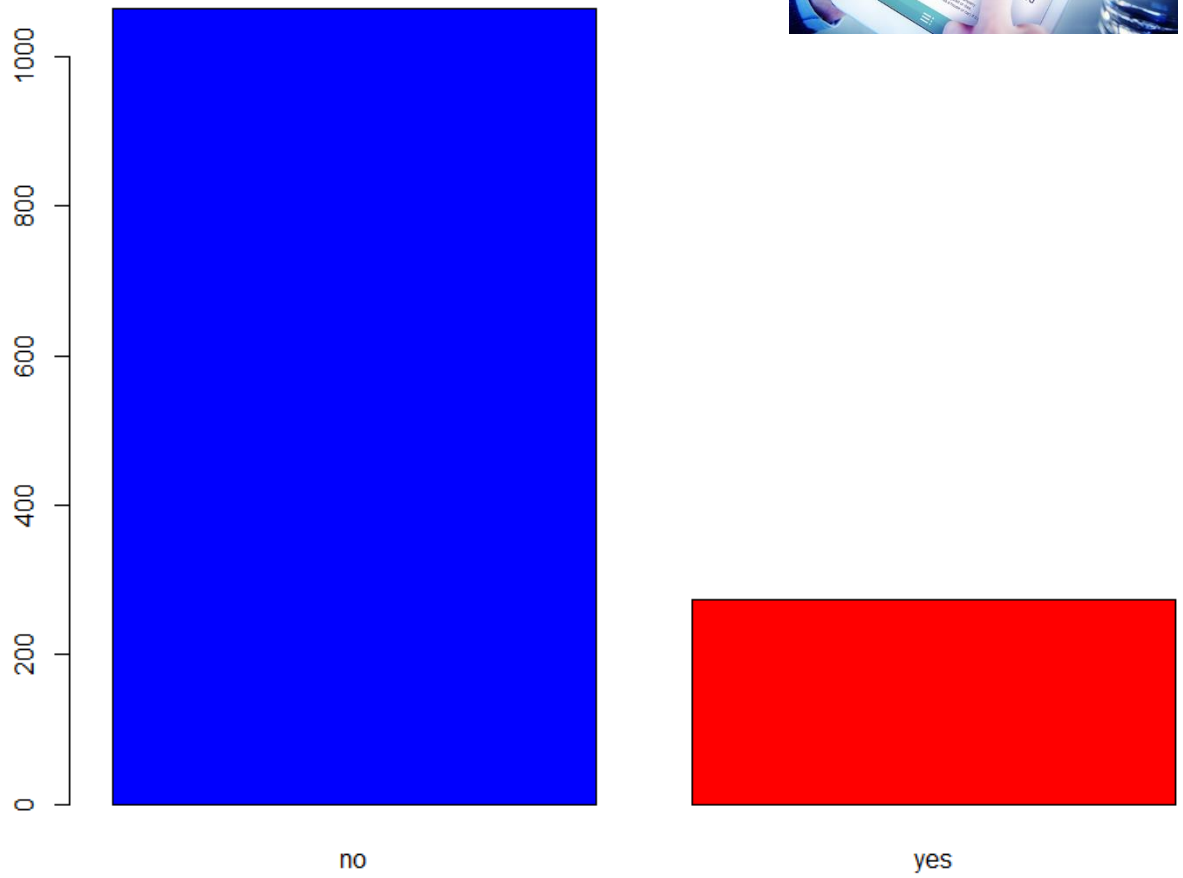
```
> c

  0   1   2   3   4   5
574 324 240 157  25  18
```

In USA also some have 5 children!

Smokers in data

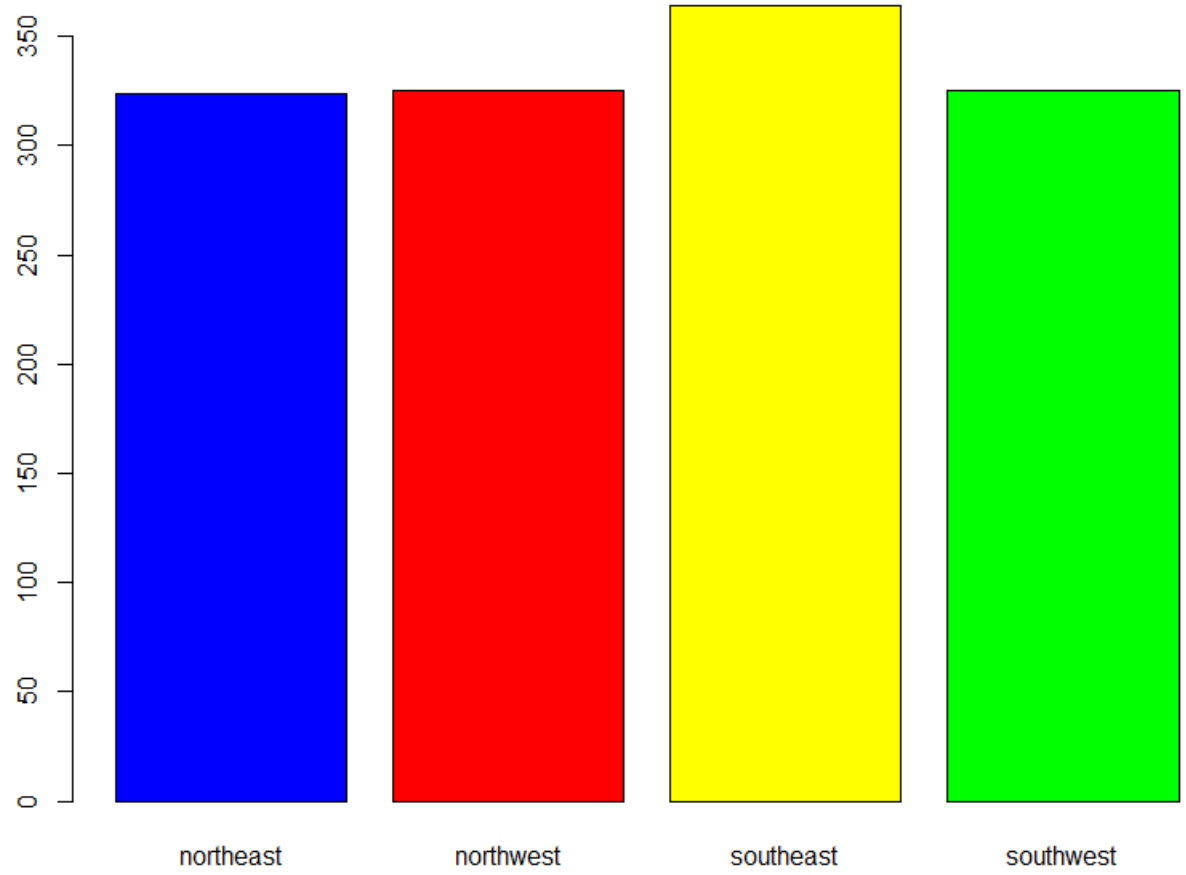

```
d = table(insurance$smoker)

d

barplot(d, col = c("blue", "red"))
```

```
> d

  no  yes
1064  274
```



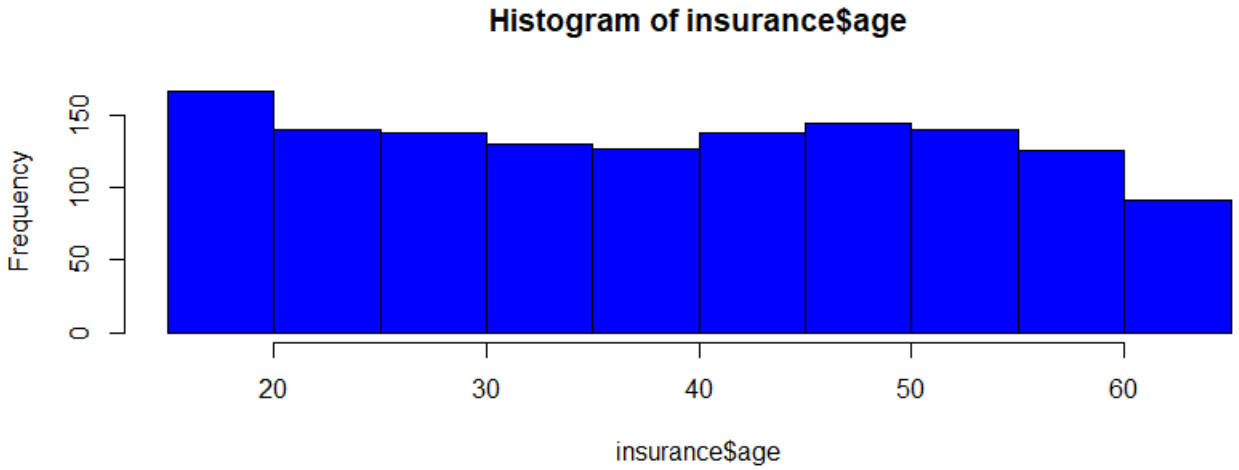
```
a = table(insurance$region)

a

barplot(a, col = c("blue", "red", "yellow", "green"))
```


```
> a

northeast northwest southeast southwest
      324       325       364       325
```

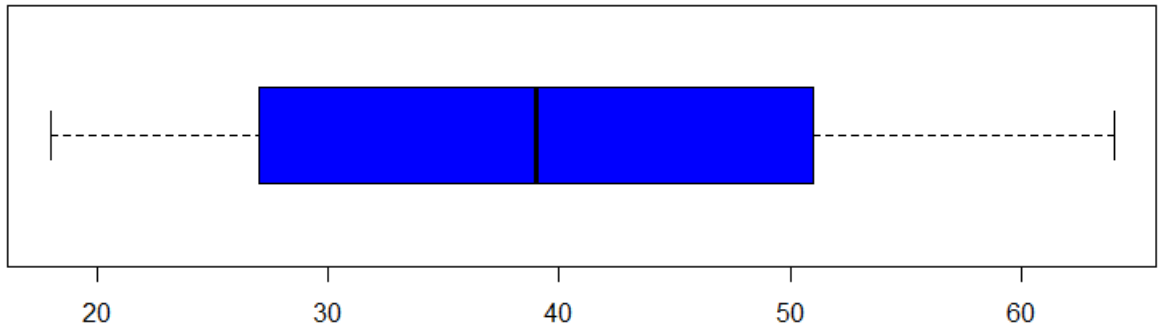


Age

```
par(mfrow = c(2,1))

hist(insurance$age, col = 'blue')
boxplot(insurance$age, col = 'blue',
        horizontal = T)

summary(insurance$age)
```

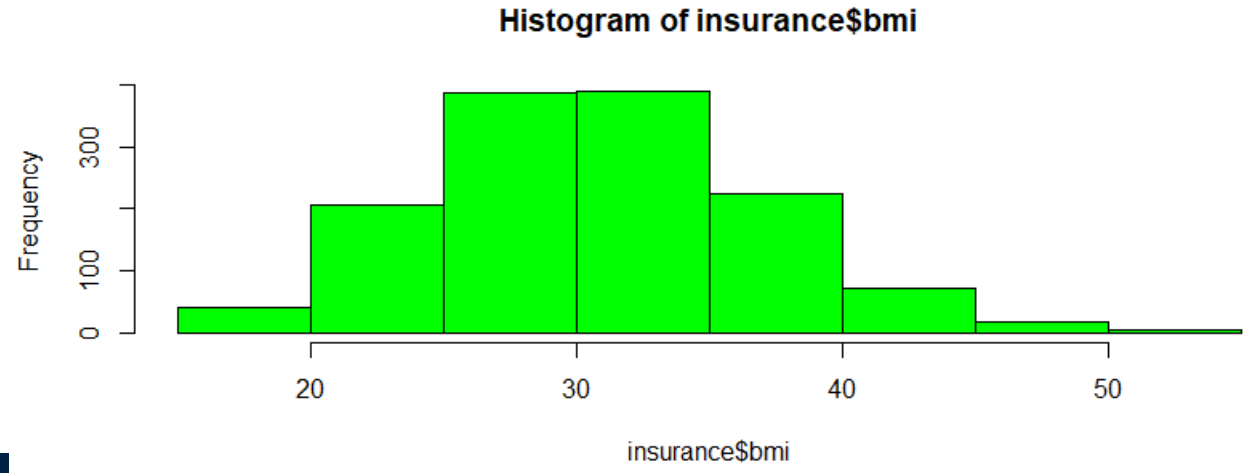



```
> summary(insurance$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   27.00   39.00   39.21   51.00   64.00
```

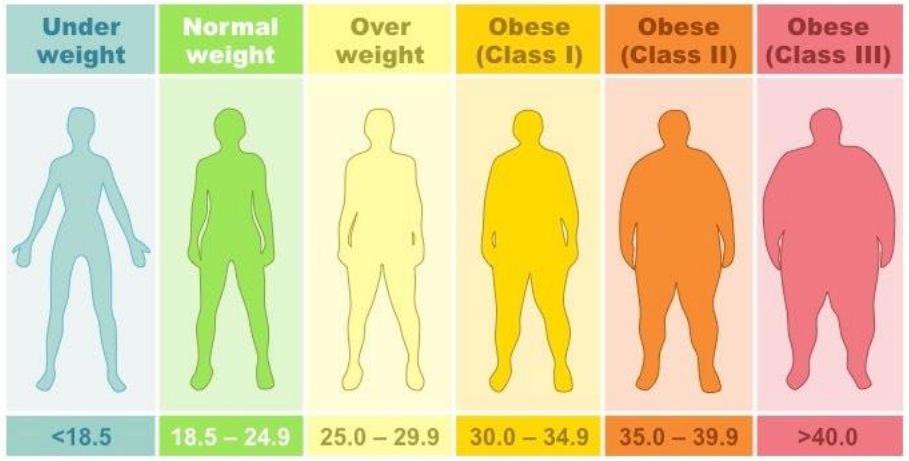


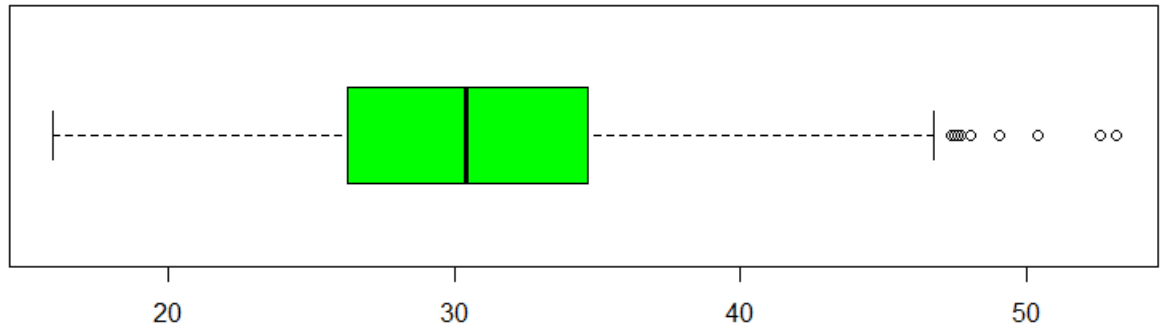
BMI

```
hist(insurance$bmi, col = 'green')
boxplot(insurance$bmi, col = 'green',
        horizontal = T)
```




```
> summary(insurance$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.96   26.30   30.40   30.66   34.69   53.13
```

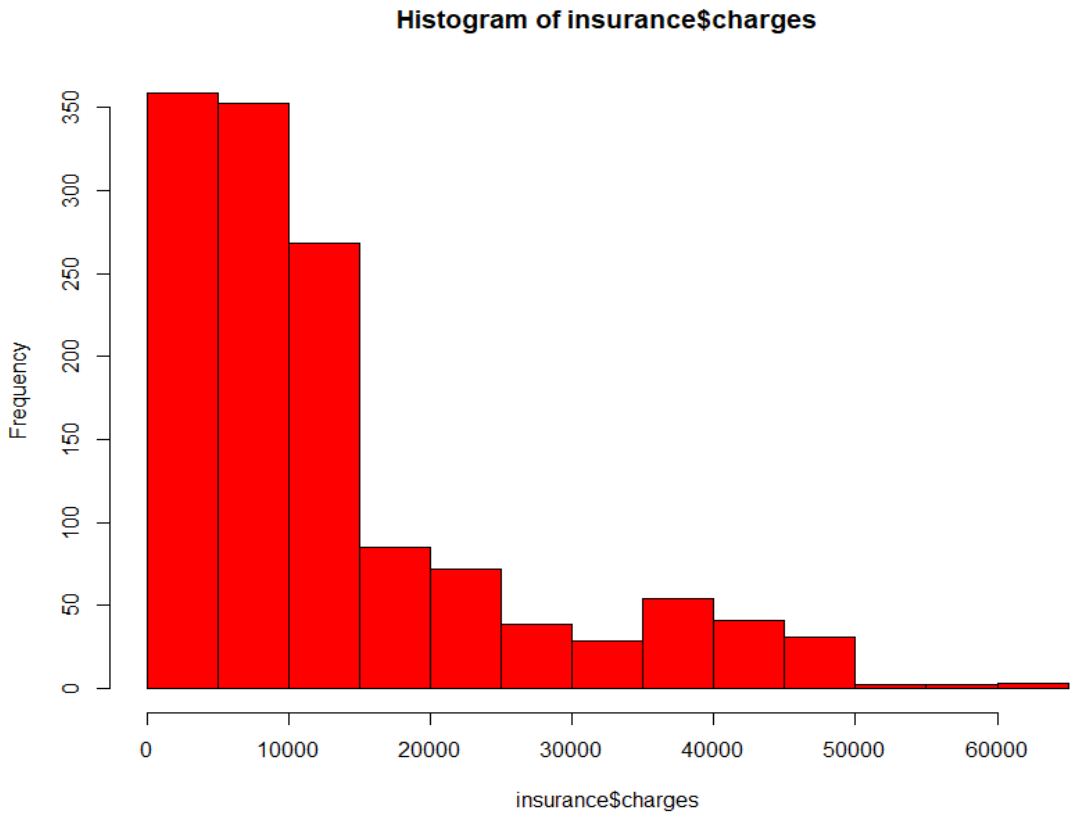


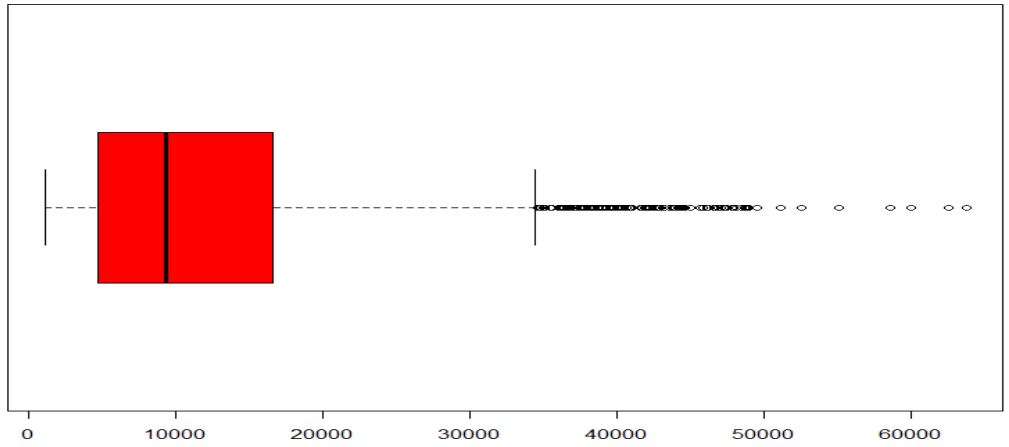


Charges

```
hist(insurance$charges, col = 'red')
boxplot(insurance$charges, col = 'red',
        horizontal = T)
```

```
> summary(insurance$charges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1122    4740    9382   13270   16640   63770
```

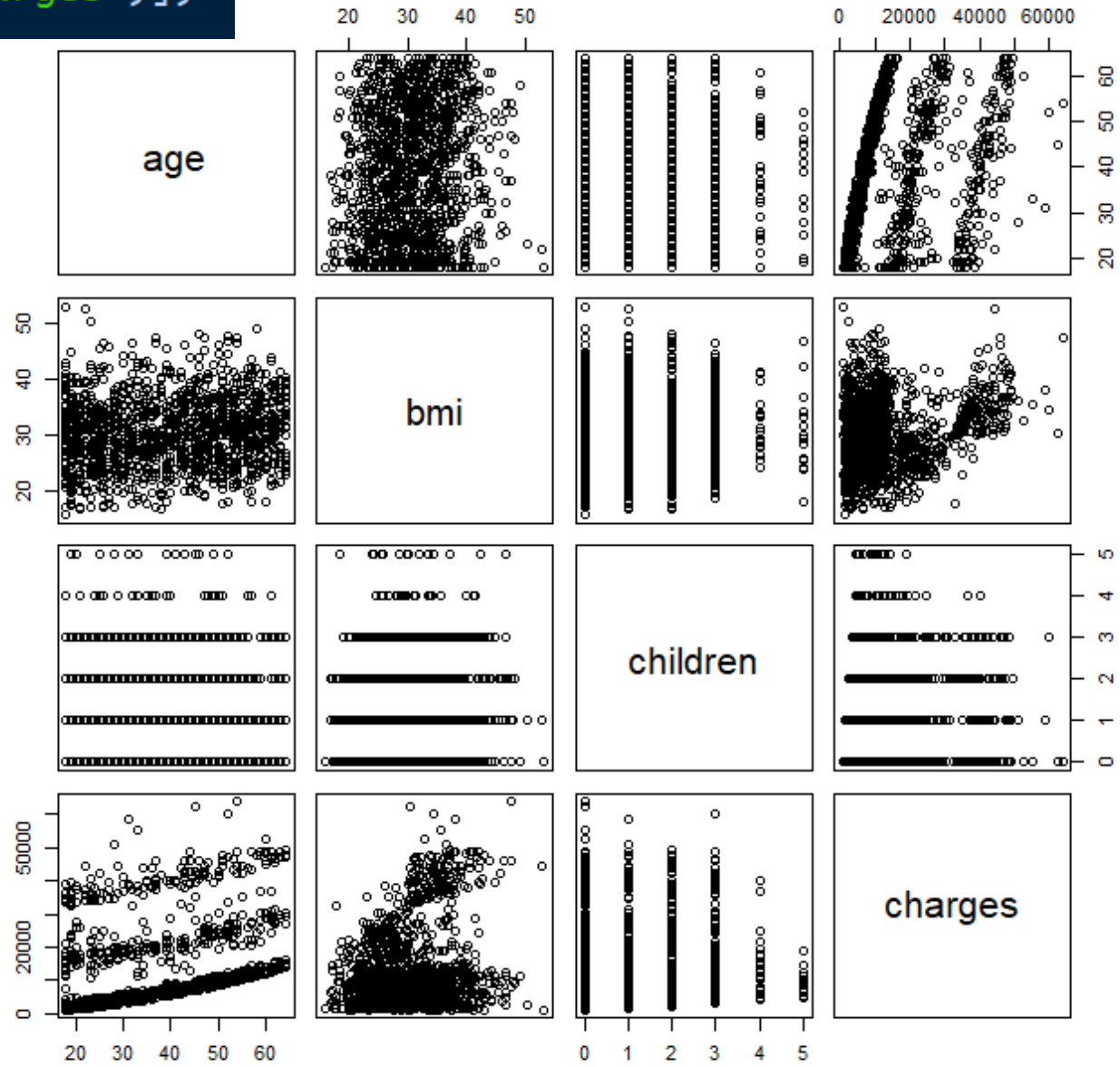




```
> cor(insurance[c("age", "bmi", "children", "charges")])
                age        bmi   children    charges
age       1.0000000 0.1092719 0.04246900 0.29900819
bmi       0.1092719 1.0000000 0.01275890 0.19834097
children  0.0424690 0.0127589 1.00000000 0.06799823
charges   0.2990082 0.1983410 0.06799823 1.00000000
```

```
pairs(insurance[c("age", "bmi", "children", "charges")])
```

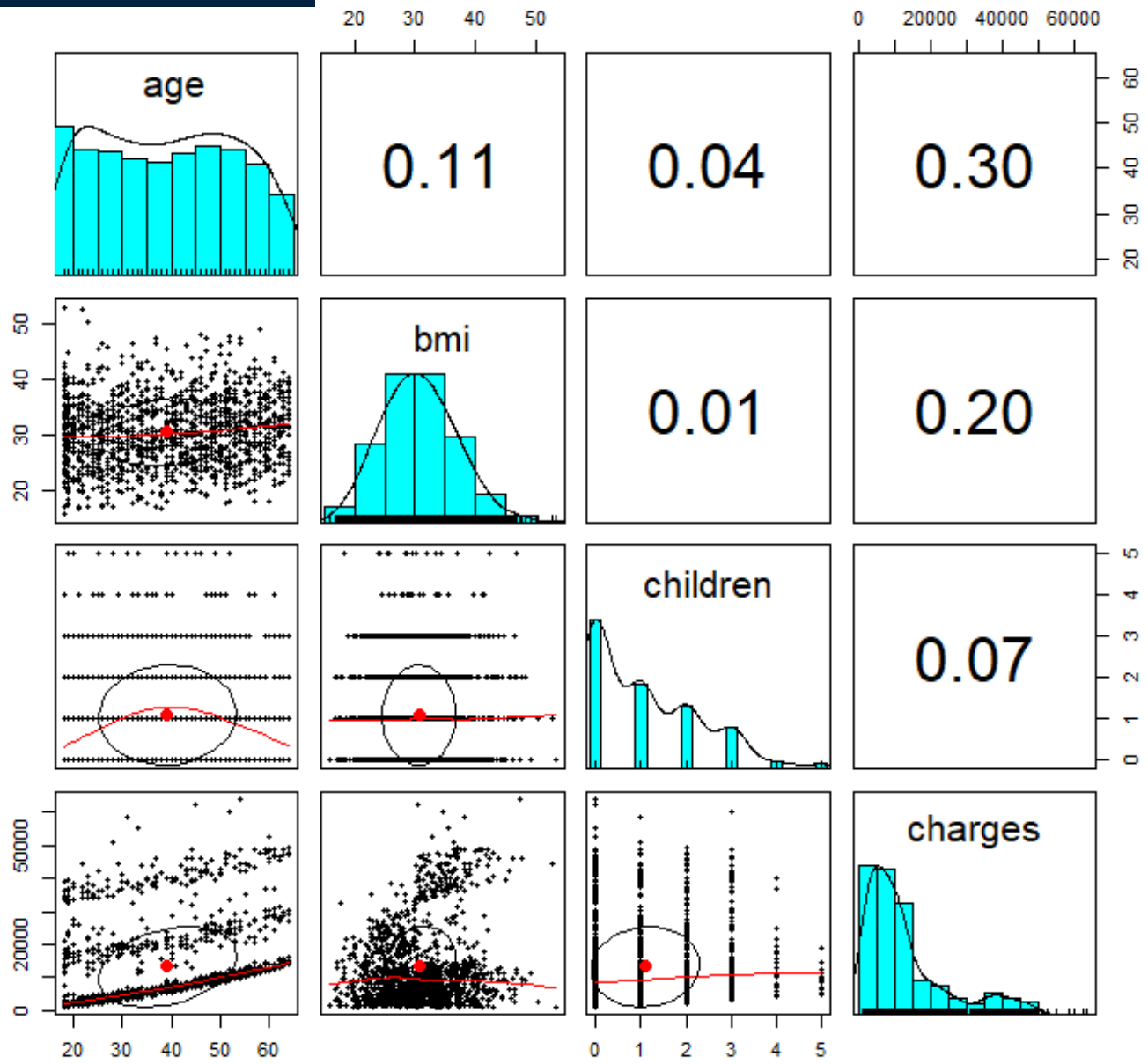

```
> cor(insurance[c("age", "bmi", "children", "charges")])
                age        bmi   children    charges
age       1.0000000 0.1092719 0.04246900 0.29900819
bmi       0.1092719 1.0000000 0.01275890 0.19834097
children  0.0424690 0.0127589 1.00000000 0.06799823
charges   0.2990082 0.1983410 0.06799823 1.00000000
```



Dr Vinod on Linear Regression 8971073111
vinodanalytics@gmail.com

```
library(psych)
pairs.panels(insurance[c("age", "bmi", "children", "charges")])
```


```
> cor(insurance[c("age", "bmi", "children", "charges")])
                age       bmi   children    charges
age       1.0000000 0.1092719 0.04246900 0.29900819
bmi       0.1092719 1.0000000 0.01275890 0.19834097
children  0.0424690 0.0127589 1.00000000 0.06799823
charges   0.2990082 0.1983410 0.06799823 1.00000000
```



Dr Vinod on Linear Regression 8971073111
vinodanalytics@gmail.com

Model

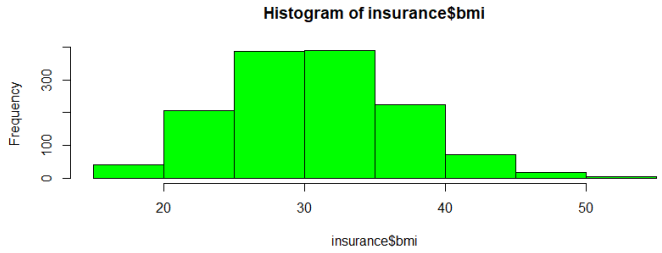
```
ins_model <- lm(charges~ age + children + bmi + sex + smoker + region,
                data = insurance)

ins_model
```

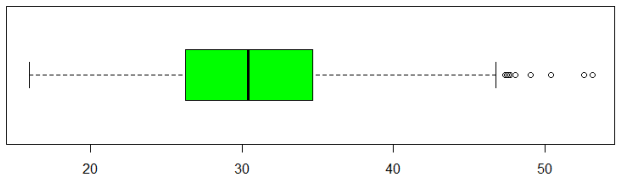

```
> ins_model

Call:
lm(formula = charges ~ age + children + bmi + sex + smoker +
    region, data = insurance)

Coefficients:
    (Intercept)              age         children
       -11938.5            256.9            475.5
            bmi          sexmale        smokeryes
          339.2           -131.3          23848.5
regionnorthwest  regionsoutheast  regionsouthwest
          -353.0           -1035.0           -960.1
```


```
> summary(ins_model)

Call:
lm(formula = charges ~ age + children + bmi + sex + smoker +
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
bmi                339.2       28.6  11.860  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Can we do better?


```
insurance$age2 <- insurance$age^2

insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
```

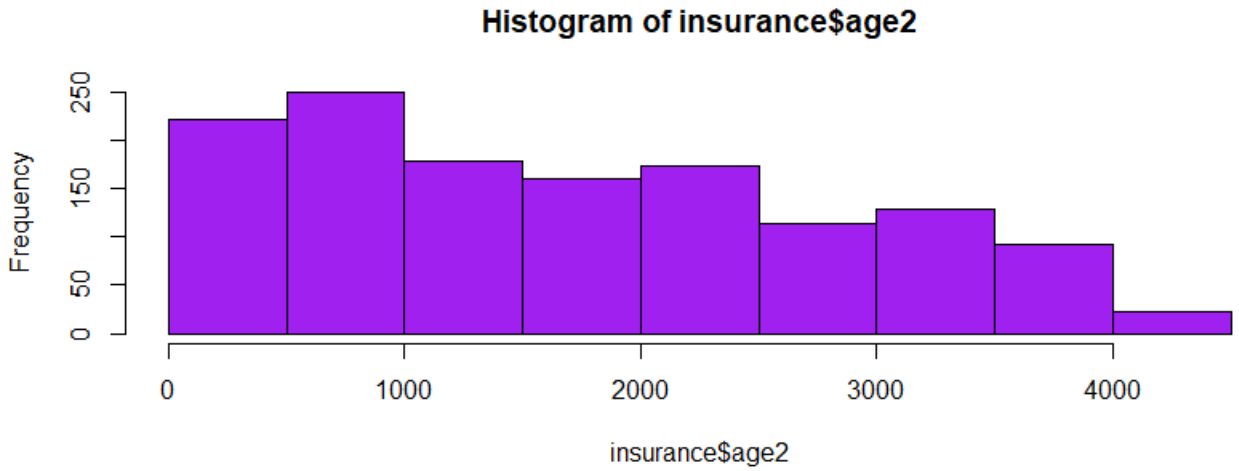


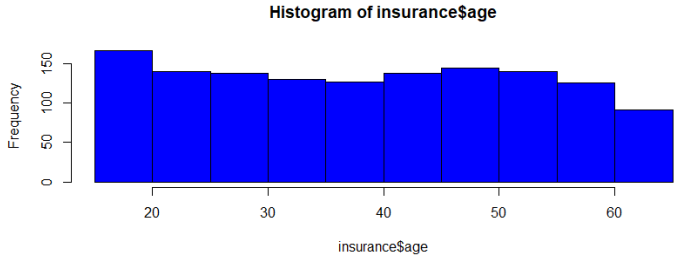


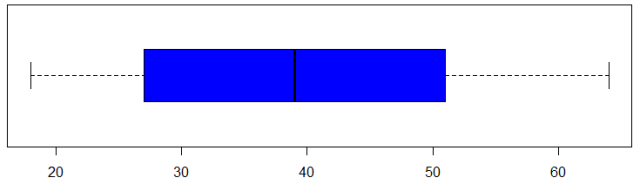
Age squared

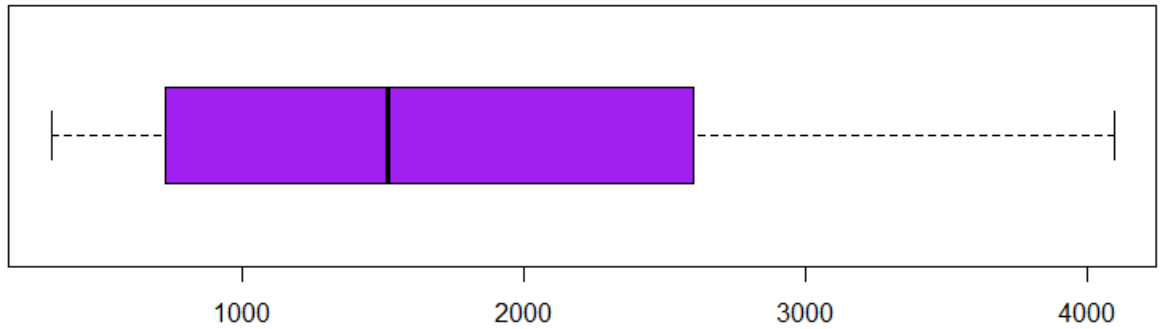
```
hist(insurance$age2, col = 'purple')

summary(insurance$age2)
```

```
> summary(insurance$age2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    324     729    1521    1734    2601    4096
```







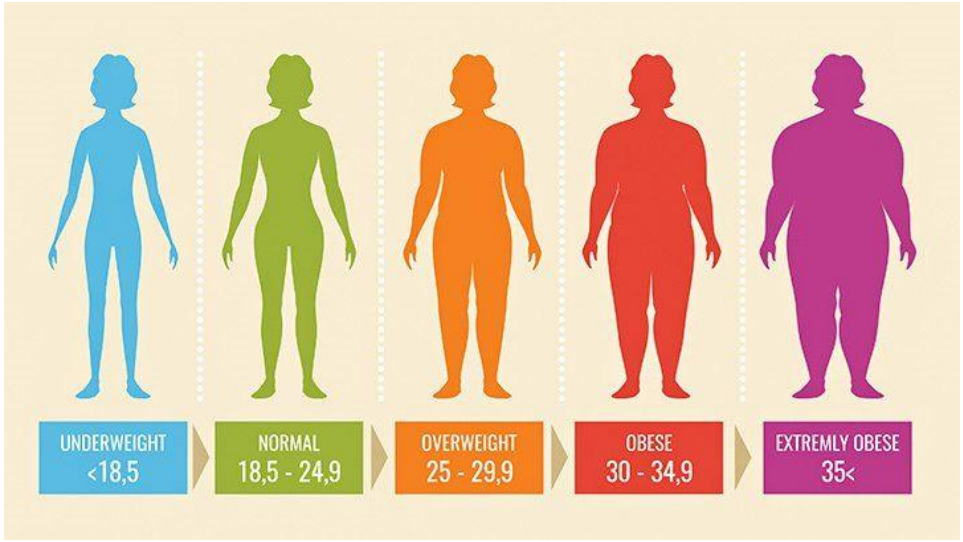


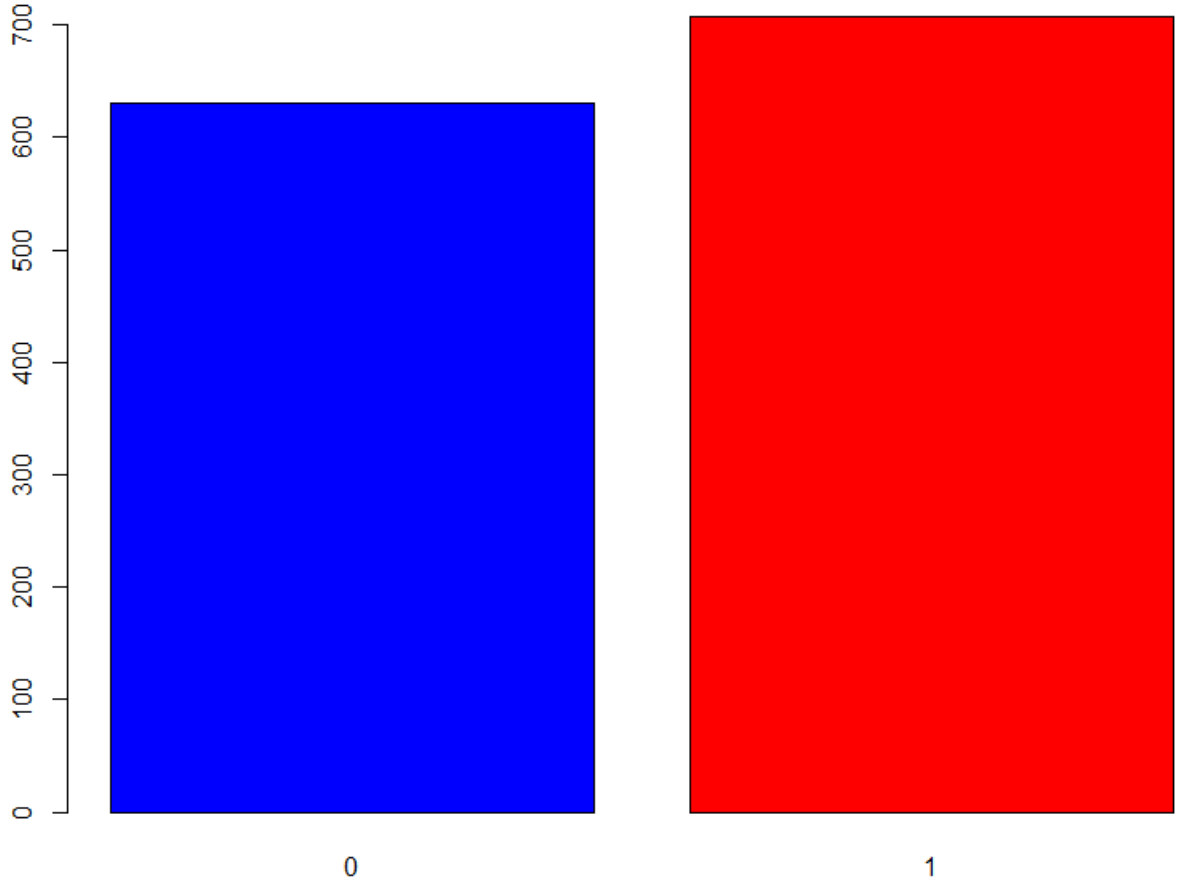

```
> summary(insurance$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   27.00   39.00   39.21   51.00   64.00
```

BMI > 30 and < 30

```
e = table(insurance$bmi30)

e

barplot(e, col = c("blue", "red"))
```

```
> e

  0   1
631 707
```





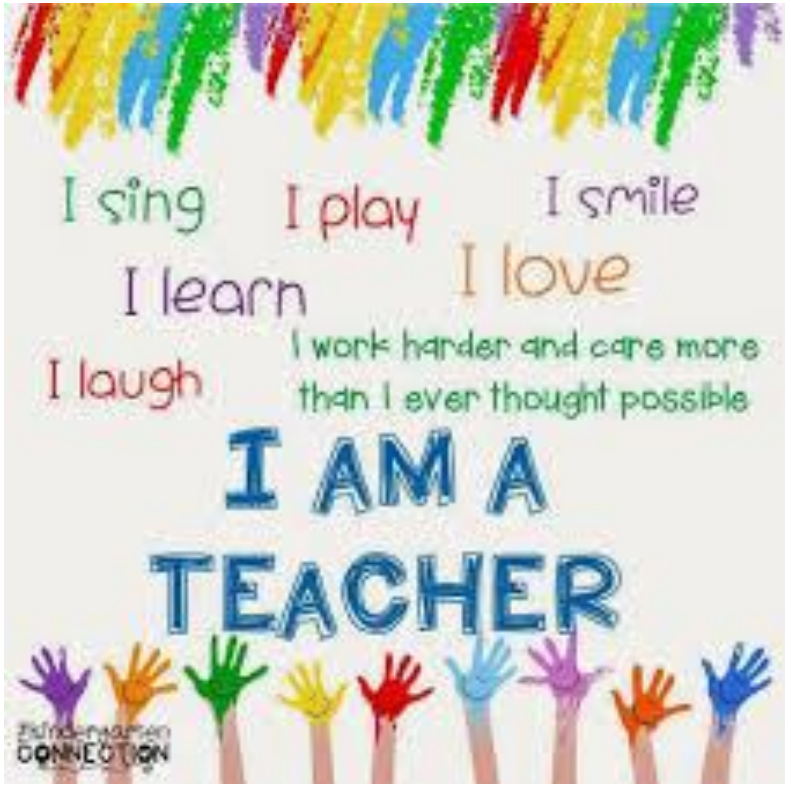
```
ins_model2 <- lm(charges ~ age + age2 + children +
                   bmi + sex + bmi30*smoker + region,
                 data = insurance)

summary(ins_model2)
```

Age = 40, Age2= 1600, 3, 20, M, 0, sw

```
> summary(ins_model)

Call:
lm(formula = charges ~ age + children + bmi + sex + smoker +
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
bmi                339.2       28.6  11.860  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```


```
> summary(ins_model2)

Call:
lm(formula = charges ~ age + age2 + children + bmi + sex + bmi
30 *
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-17296.4  -1656.0  -1263.3   -722.1  24160.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       134.2509  1362.7511   0.099 0.921539
age               -32.6851    59.8242  -0.546 0.584915
age2                3.7316     0.7463   5.000 6.50e-07 ***
children          678.5612   105.8831   6.409 2.04e-10 ***
bmi               120.0196    34.2660   3.503 0.000476 ***
sexmale          -496.8245   244.3659  -2.033 0.042240 *
bmi30           -1000.1403   422.8402  -2.365 0.018159 *
smokeryes       13404.6866   439.9491  30.469  < 2e-16 ***
regionnorthwest  -279.2038   349.2746  -0.799 0.424212
regionsoutheast  -828.5467   351.6352  -2.356 0.018604 *
regionsouthwest -1222.6437   350.5285  -3.488 0.000503 ***
bmi30:smokeryes 19810.7533   604.6567  32.764  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared:  0.8664,    Adjusted R-squared:  0.8653
F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Happy Learning!



When adding a dummy variable to a regression model, one category is always left out to serve as the reference category. The estimates are then interpreted relative to the reference. In our model, R automatically held out the `sexfemale`, `smokerno`, and `regionnortheast` variables, making female non-smokers in the northeast region the reference group. Thus, males have $131.40 less medical expenses each year relative to females and smokers cost an average of $23,847.50 more than non-smokers per year. The coefficient for each of the three regions in the model is negative, which implies that the reference group, the northeast region, tends to have the highest average expenses.

By default, R uses the first level of the factor variable as the reference. If you would prefer to use another level, the `relevel()` function can be used to specify the reference group manually. Use the `?relevel` command in R for more information.

The results of the linear regression model make logical sense: old age, smoking, and obesity tend to be linked to additional health issues, while additional family member dependents may result in an increase in physician visits and preventive care such as vaccinations and yearly physical exams. However, we currently have no sense of how well the model is fitting the data. We'll answer this question in the next section.